Microphone-array ego-noise reduction algorithms for auditory micro aerial vehicles

Lin Wang, Andrea Cavallaro

***Abstract*—When a micro aerial vehicle (MAV) captures sounds emitted by a ground or aerial source, its motors and propellers are much closer to the microphone(s) than the sound source, thus leading to extremely low signal-to-noise ratios (SNR), *e.g.* -15 dB. While microphone-array techniques have been investigated intensively, their application to MAV-based ego-noise reduction has been rarely reported in the literature. To fill this gap, we implement and compare three types of microphone-array algorithms to enhance the target sound captured by an MAV. These algorithms include a newly emerged technique, time-frequency spatial filtering, and two well-known techniques, beamforming and blind source separation. In particular, based on the observation that the target sound and the ego-noise usually have concentrated energy at sparsely isolated time-frequency bins, we propose to use the time-frequency processing approach, which formulates a spatial filter that can enhance a target direction based on local direction of arrival estimates at individual time-frequency bins. By exploiting the time-frequency sparsity of the acoustic signal, this spatial filter works robustly for sound enhancement in the presence of strong ego-noise. We analyze in details the three techniques and conduct a comparative evaluation with real-recorded MAV sounds. Experimental results show the superiority of blind source separation and time-frequency filtering in low-SNR scenarios.**

***Index Terms*—Acoustic sensing, ego-noise reduction, micro aerial vehicles, microphone array**

I. Introduction

With the ability of hovering above the terrain and moving in 3D, multi-rotor micro aerial vehicles (MAV) are an ideal mobile sensing platform that can be equipped with cameras, laser scanners, ultrasonic radars and microphones [1]. While visual sensing has already attracted considerable attention for search and rescue operations, personal and professional video capturing [2]–[6], acoustic sensing using microphones mounted on the MAV is a new and emerging topic. When deploying MAVs in search and rescue operations, acoustic sensing is desirable in order to detect sound-emitting targets especially with low visibility or visual obstacles (*e.g.* in case of a victim underneath debris) [7]–[12]. Moreover, MAVs for multimedia broadcasting could stream both audio and video signals to remote terminals [13], [14].

The main obstacle for effective MAV-based acoustic sensing is the strong ego-noise generated by motors and propellers [15], which masks the target sounds and degrades the recorded signal significantly [16]. Microphone-array ego-noise reduction techniques are needed to enhance the target sound. Since the motors and propellers are closer to the microphones than the target sound source, the MAV sound recording usually presents an extremely low signal-to-noise ratio (SNR), which considerably degrades the performance of most microphone-array signal processing algorithms. The spectrum of this nonstationary ego-noise depends on the rotation speed of each motor, which changes over time. Moreover, the microphones move with the MAV thus leading to a dynamic acoustic mixing network. Finally, natural and motion-induced wind increases the noise components captured by the microphones.

While microphone-array techniques have been investigated intensively in the last decades [17]–[20], most algorithms were developed for indoor speech processing. Moreover, the application to extremely low-SNR scenarios (*e.g.* <-15 dB) has been rarely reported [21], [22]. In addition to this, only a few works have specifically addressed the challenging MAV-based ego-noise problem [13], [16], [23]–[27]. These works can be categorized as supervised and unsupervised approaches.

Supervised approaches need additional sensors to estimate or to predict the ego-noise. Two types of supervised methods have been proposed for ego-noise reduction, namely template-based [23] and reference-based [13], [26], [27]. *Template-based methods* build a noise template database from which the spectrum [23] or the correlation matrix [28] of the ego-noise can be estimated corresponding to the motor rotation speed and the MAV behaviour. The estimated ego-noise information can be used to design single-channel spectral filters [23] or multichannel adaptive beamformer [29], [30] for ego-noise reduction, and can also be applied to noise-robust source localization [28]. To avoid using monitoring sensors, non-negative matrix factorization can be employed to learn noise bases from pre-recorded training data and then to estimate the noise spectrum online from the noisy recording. While this approach has already been applied to ground robots [31], [32], its performance for MAV ego-noise, which is nonstationary and stronger, has not been reported yet. *Reference-based methods* use reference microphones installed close to the propellers to pick up motor noises and then cancel them out with an adaptive filter [13], [26], [27]. Insulation materials are necessary to prevent the reference microphones from picking up the target sound. Overall, the need for dedicated monitoring sensors limits the versatility of supervised approaches.

Unsupervised approaches perform ego-noise reduction using only the microphone signals. To date, only two types of unsupervised approaches have been applied to MAVs, namely fixed beamforming [24], [25] and blind source separation [16].

Manuscript received: February 12, 2017

The authors are with Centre for Intelligent Sensing, Queen Mary University of London, London, UK (e-mail: lin.wang@qmul.ac.uk; a.cavallaro@qmul.ac.uk)

This work was supported in part by the ARTEMIS-JU and the UK Technology Strategy Board (Innovate UK) through the COPCAMS Project, under grant 332913.

Delay-and-sum (fixed) *beamforming* enhances the sound from a desired location by coherently delaying and summing multichannel microphone signals. Fixed beamforming relies only on the array geometry and the target sound location, and is robust to low SNRs and MAV movement. However, to obtain satisfactory noise reduction performance, this approach usually requires a large-size array, *e.g.* 16 microphones in an octagonal array with a diameter of around 2 m [24], [25]. *Blind source separation* (BSS) recovers unknown source signals from the observed mixture by blindly estimating a demixing network [33]. BSS suppresses the ego-noise without knowing the locations of the microphones and the target sound source [16]. However, the performance of BSS degrades in a dynamic scenario with a moving MAV.

To fill the gap between the extensive work in microphone-array signal processing and the new applications to MAVs, after introducing the problem formulation in Section II, we present in Section III three types of unsupervised microphone-array algorithms that can be applied for ego-noise reduction: time-frequency spatial filtering (a recently emerged technique), beamforming and blind source separation. The time-frequency processing approach was originally proposed for indoor speech processing, which formulates a spatial filter by exploiting the time-frequency sparsity of speech signals [34], [35]. Based on the observation that the captured acoustic signals usually have sparsely concentrated energy in the time-frequency domain, we propose to apply this technique to MAV-based sound processing. Moreover, we build a hardware prototype to test and compare the algorithms, as discussed in Section IV. Finally, Section V draws conclusions.

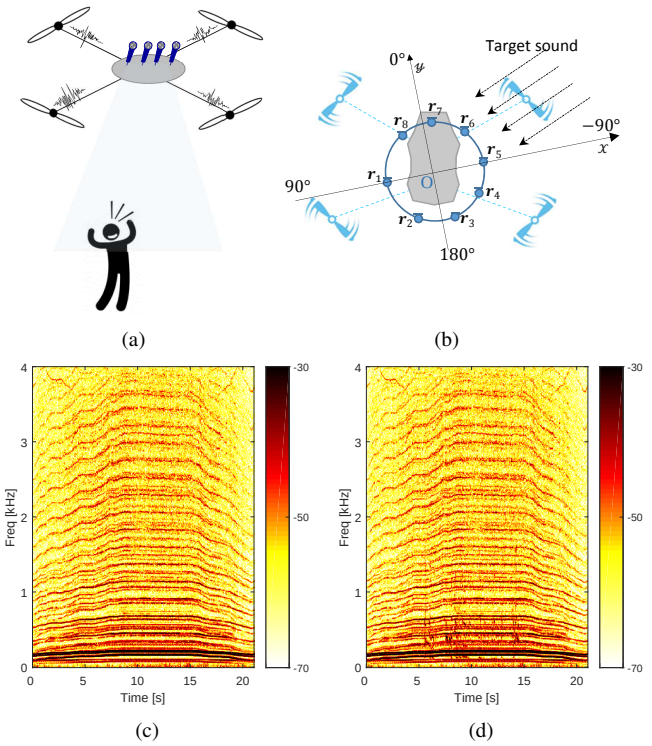


Fig. 1. (a) A hovering multi-rotor MAV with a microphone array capturing a target sound. (b) Geometrical configuration for microphone-array signal processing. (c) Time-frequency spectrum of the ego-noise recorded from an operational MAV. The ego-noise consists of harmonic components and broadband noise. (d) Time-frequency spectrum of the noisy recording composed of ego-noise and a target sound (speech), which occurs during 5-15 s with SNR -15 dB. The target sound is almost fully masked by the background noise.

II. Problem Formulation

Let a circular array with M microphones mounted on a multi-rotor MAV capture the sound emitted by a target source (Fig. 1(a)). The locations of the microphones in a 2D coordinate system, as shown in Fig. 1(b), are $\boldsymbol{R} = [\boldsymbol{r}_1, \cdots, \boldsymbol{r}_M]$, where $\boldsymbol{r}_m = [r_{mx}, r_{my}]^{\mathrm{T}}$ is the location of the m-th microphone and the superscript $(\cdot)^{\mathrm{T}}$ denotes the transpose. The target sound source is in the far field and emits sound with direction of arrival (DOA) θ_d. The microphone signal, $\boldsymbol{x}(n) = [x_1(n), \cdots, x_M(n)]^{\mathrm{T}}$, contains both the target sound, $\boldsymbol{s}(n) = [s_1(n), \cdots, s_M(n)]^{\mathrm{T}}$, and the ego-noise, $\boldsymbol{v}(n) = [v_1(n), \cdots, v_M(n)]^{\mathrm{T}}$, *i.e.*

$$\boldsymbol{x}(n) = \boldsymbol{s}(n) + \boldsymbol{v}(n), \tag{1}$$

or, written in the short-time Fourier transform (STFT) domain:

$$\boldsymbol{x}(k,l) = \boldsymbol{s}(k,l) + \boldsymbol{v}(k,l), \tag{2}$$

where k and l are the frequency and frame indices, respectively. Let K and L be the total number of frequency bins and time frames, respectively.

Given $\boldsymbol{x}(n)$, $\boldsymbol{R}$ and θ_d, we aim to design a spatial filter $\boldsymbol{w}(k,l) = [w_1(k,l), \cdots, w_M(k,l)]^{\mathrm{T}}$ that extracts the target sound from the noisy recording via

$$y(k,l) = \boldsymbol{w}^{\mathrm{H}}(k,l)\boldsymbol{x}(k,l), \tag{3}$$

where the superscript $(\cdot)^{\mathrm{H}}$ denotes the Hermitian transpose.

Fig. 1(c) shows an example of the ego-noise recorded from an operational MAV. The ego-noise consists mainly of narrow-band harmonic components and broadband noise. The *harmonic* noise is the mechanical sound generated by the rotating motors, with energy peaks at isolated frequency bins. The *braodband* noise is generated by the rotating propellers cutting the air, with its energy spreading uniformly throughout the frequency spectrum. The fundamental frequency (pitch) of the harmonics usually varies with the motor rotation speed [23], leading to nonstationarity of the ego-noise. For instance, the ego-noise shown in Fig. 1(c) was produced when the motor speed increased monotonically in the first 7.5 s (seconds), then remained stable during 7.5-15 s, and finally decreased in the last 6 s. The pitch of the harmonics varies similarly to the motor speed. We modelled the ego-noise in [16] as the sum of multiple directional point-source noises, which show high correlation at harmonic frequencies, plus one diffuse noise, which shows low correlation at high frequencies but high correlation at low frequencies. Microphone-array techniques, which exploit the correlation of the acoustic signal among microphones, are thus suitable to address this ego-noise reduction problem. Fig. 1(d) shows an example of noisy recording with a target sound (speech) present during 5-15 s with SNR -15 dB. Comparing Fig. 1(c) and Fig. 1(d), only minor differences can be observed as the target sound is masked by the background noise and is almost invisible.

In the following section we discuss microphone-array algorithms that are appropriate for MAV-based noise reduction, assuming that the MAV hovers stably while recording the sound from a static source (*i.e.* the locations of the microphones and the sound source are fixed). We assume a low-reverberant environment, as MAVs are mainly deployed outdoors, and we do not take into account the noise produced by natural wind.

III. Microphone-array algorithms for MAVs

A. Beamforming

Beamforming is a widely used microphone-array technique for directional sound acquisition [17]. A *fixed beamformer* enhances the sound from a specific direction by coherently delaying and summing the signals from multiple microphones based on the transmitting delays from the sound source to the microphones:

$$y_{\text{BF}}(k,l,\theta_d) = \frac{1}{M}\sum_{m=1}^{M} x(k,l)\mathrm{e}^{j2\pi f_k \tau(1,m,\theta_d)}, \tag{4}$$

where f_k denotes the frequency at the k-th bin, $\tau(m_1,m_2,\theta_d) = \frac{\|\boldsymbol{r}_{m_2}-\boldsymbol{r}_{\theta_d}\|-\|\boldsymbol{r}_{m_1}-\boldsymbol{r}_{\theta_d}\|}{c}$ is the delay between two microphones m_1 and m_2 with respect to the sound coming from θ_d, $\boldsymbol{r}_{\theta_d}$ is the location of the far-field sound source with DOA θ_d, and c is the velocity of sound. The performance of a fixed beamformer is mainly determined by the geometrical configuration of the array and the sound source (*i.e.* $\boldsymbol{R}$ and θ_d), and usually is not related to the acoustic signals received by the microphones.

Adaptive beamformers analyze statistical characteristics of the microphone signal to enhance the target sound without knowing the locations of the microphones and the target sound source. Adaptive beamformers suppress noise more efficiently than (delay-and-sum) fixed beamformers. Several criteria can be applied in the design of an adaptive beamformer, such as minimum variance distortionless response (MVDR), maximum speech-to-noise ratio (MaxSNR), and multichannel Wiener filter (MWF) [36]–[38]. These adaptive beamformers typically require the knowledge of the correlation matrix of the target sound or of the noise signal. The microphone signal in the time-frequency domain $\boldsymbol{x}(k,l) = \boldsymbol{s}(k,l) + \boldsymbol{v}(k,l)$, the correlation matrices of the microphone signal $\boldsymbol{\Phi}_{xx}(k,l)$, the target signal $\boldsymbol{\Phi}_{ss}(k,l)$, and the noise signal $\boldsymbol{\Phi}_{vv}(k,l)$ are, respectively, defined as

$$\boldsymbol{\Phi}_{xx}(k,l) = \mathrm{E}\{\boldsymbol{x}(k,l)\boldsymbol{x}^{\mathrm{H}}(k,l)\}, \tag{5}$$

$$\boldsymbol{\Phi}_{ss}(k,l) = \mathrm{E}\{\boldsymbol{s}(k,l)\boldsymbol{s}^{\mathrm{H}}(k,l)\}, \tag{6}$$

$$\boldsymbol{\Phi}_{vv}(k,l) = \mathrm{E}\{\boldsymbol{v}(k,l)\boldsymbol{v}^{\mathrm{H}}(k,l)\}, \tag{7}$$

where $\mathrm{E}\{\cdot\}$ denotes the mathematical expectation. Assuming that the target and the noise signals are statistically independent, it follows that

$$\boldsymbol{\Phi}_{xx}(k,l) = \boldsymbol{\Phi}_{ss}(k,l) + \boldsymbol{\Phi}_{vv}(k,l). \tag{8}$$

If the correlation matrices in (8) are known, the adaptive beamformer can be formulated easily. For instance, applying generalized eigen-vector decomposition (GEVD) to a matrix pair ($\boldsymbol{\Phi}_{xx}(k,l)$, $\boldsymbol{\Phi}_{vv}(k,l)$), it follows that [38]

$$\boldsymbol{\Phi}_{xx}(k,l)\boldsymbol{E}(k,l) = \boldsymbol{\Phi}_{vv}(k,l)\boldsymbol{E}(k,l)\boldsymbol{\Lambda}(k,l), \tag{9}$$

where $\boldsymbol{\Lambda}(k,l)$ is a diagonal matrix containing M generalized eigen-values $\lambda_1(k,l) \geq \cdots \geq \lambda_M(k,l)$, $\boldsymbol{E}(k,l) = [\boldsymbol{e}_1(k,l),\cdots,\boldsymbol{e}_M(k,l)]$ consists of M generalized eigen-vectors corresponding to $\lambda_1(k,l),\cdots,\lambda_M(k,l)$, respectively. The MaxSNR beamformer is defined as the generalized eigen-vector corresponding to the largest eigen-value, *i.e.*

$$\boldsymbol{w}_{\text{MaxSNR}}(k,l) = \boldsymbol{e}_1(k,l), \tag{10}$$

with the output being

$$y_{\text{MaxSNR}}(k,l) = \boldsymbol{w}^{\mathrm{H}}_{\text{MaxSNR}}(k,l)\boldsymbol{x}(k,l). \tag{11}$$

A crucial problem in adaptive beamformer design is the estimation of the correlation matrices.

The *correlation matrix* of the microphone signal, as defined in (5), can be estimated directly by using

$$\boldsymbol{\Phi}_{xx}(k,l) = \frac{1}{L}\sum_{l=1}^{L}\boldsymbol{x}(k,l)\boldsymbol{x}^{\mathrm{H}}(k,l). \tag{12}$$

Estimating the correlation matrix of the MAV ego-noise is a challenging task. If the noise signal is known for the whole duration of the signal, the noise correlation matrix, as defined in (7), can be estimated similarly to (12), *i.e.*

$$\boldsymbol{\Phi}^{\text{Ideal}}_{vv}(k,l) = \frac{1}{L}\sum_{l=1}^{L}\boldsymbol{v}(k,l)\boldsymbol{v}^{\mathrm{H}}(k,l). \tag{13}$$

This scheme works only in an ideal situation. An alternative scheme is to estimate the noise correlation matrix using the microphone signal in noise-only periods $\mathbb{L}_v$, *i.e.*

$$\boldsymbol{\Phi}^{\text{vad}}_{vv}(k,l) = \frac{1}{L_v}\sum_{l\in\mathbb{L}_v}\boldsymbol{x}(k,l)\boldsymbol{x}^{\mathrm{H}}(k,l), \tag{14}$$

where L_v is the total number of frames in $\mathbb{L}_v$. This scheme is widely used in speech processing [38] but not suitable for MAV sound recording because it is difficult to find a voice activity detector (VAD) that can reliably detect the noise-only period especially when the ego-noise is nonstationary and the SNR is extremely low.

Two other ego-noise correlation matrix estimation schemes were proposed in the works which applied the GEVD-MUSIC (MUltiple SIgnal Classification) algorithm to MAV-based source localization [7], [28], [39]. The first scheme simply uses an identity matrix as the estimate of the noise correlation matrix [28], *i.e.*

$$\boldsymbol{\Phi}^{\text{identity}}_{vv}(k,l) = \boldsymbol{I}_M, \tag{15}$$

where $\boldsymbol{I}_M$ denotes an $M\times M$ identity matrix. This scheme assumes the noise signals received at multiple microphones to be uncorrelated with each other, which is not the case for MAV ego-noise. The second scheme incrementally estimates the noise correlation matrix, assuming that the L_T frames preceding the current frame contain only noise [7], [39], *i.e.*

$$\boldsymbol{\Phi}^{\text{inc}}_{vv}(k,l) = \frac{1}{L_T}\sum_{l'=l-L_T}^{l-1}\boldsymbol{x}(k,l')\boldsymbol{x}^{\mathrm{H}}(k,l'). \tag{16}$$

Obviously, this assumption does not stand when the previous L_T frames contain the target sound signal.

B. Blind source separation (BSS)

BSS performs sound enhancement by treating the target and noise signals equally and by separating all the individual sources from the mixed signals captured by the array of microphones [33]. The application of BSS to MAV-based ego-noise reduction is straightforward as the locations of the microphones and the target source are not needed [16].

BSS consists of two key components: independent component analysis (ICA) and permutation alignment [41]. ICA, which is applied per frequency bin, exploits the statistical independence between source signals to estimate a demixing matrix [42]. This demixing matrix can be interpreted as the inverse of the acoustic mixing network and can recover the source signals up to permutation ambiguities: each source can be extracted individually from the observed mixture but with a random order in the output channels. A subsequent permutation alignment procedure is needed to group the individual signals that belong to the same source so that the separated frequency-domain signals can be correctly transformed back to the time domain [40], [41].

Since we have M microphones, we apply an $M \times M$ ICA directly to the M-channel input, assuming an $M \times M$ mixing network with M independent sources, *i.e.* a target sound source component $\tilde{s}$ and $M' = M - 1$ unknown ego-noise components $\tilde{v}_1, \cdots, \tilde{v}_{M'}$. The M-channel microphone input can thus be written in the time-frequency domain as

$$\boldsymbol{x}(k,l) = \boldsymbol{H}(k,l)\boldsymbol{u}(k,l), \tag{17}$$

where $\boldsymbol{u}(k,l) = [\tilde{s}(k,l), \tilde{v}_1(k,l), \cdots, \tilde{v}_{M'}(k,l)]^{\mathrm{T}}$ is a vector containing the M sources, and $\boldsymbol{H}(k,l)$ is the $M \times M$ mixing matrix between the M sources and M microphones.

After ICA and permutation alignment, we represent the obtained demixing matrix as $\boldsymbol{W}_{\text{BSS}}(k,l)$, which approximates the inverse of the demixing matrix, *i.e.* $\boldsymbol{W}_{\text{BSS}}(k,l) \approx \boldsymbol{H}^{-1}(k,l)$. The demixed signal is obtained as

$$\boldsymbol{y}_{\text{BSS}}(k,l) = \boldsymbol{W}_{\text{BSS}}(k,l)\boldsymbol{x}(k,l) \approx \boldsymbol{u}(k,l), \tag{18}$$

where $\boldsymbol{y}_{\text{BSS}}(k,l) = [y_1(k,l), \cdots, y_M(k,l)]^{\mathrm{T}}$ is a vector of M elements, one of which is the target sound.

C. Time-frequency processing

Time-frequency (T-F) processing has emerged recently as a class of approaches that exploit the time-frequency sparsity of audio signals [34], [35]. These approaches estimate the DOA of the sound at each time-frequency bin and then combine the localization results from individual time-frequency bins for noise reduction. While the idea of local DOA-based spatial filtering was originally proposed for indoor speech processing [34], it is also suitable for MAV sound processing as the harmonic components of the ego-noise have concentrated energy peaks at isolated harmonic frequencies [16]. Likewise, target sounds such as human speech or emergency whistles also consist mainly of harmonic components.

Given the microphone signal $\boldsymbol{x}(k,l)$ and microphone location $\boldsymbol{R}$, the DOA of the sound at each time-frequency bin can be estimated by building a spatial likelihood function [34]

$$\gamma_{\text{TF}}(k,l,\theta) = \Re\left\{ \sum_{\substack{m_1,m_2=1 \\ m_1 \neq m_2}}^{M} \frac{x_{m_1}(k,l)x_{m_2}^*(k,l)}{|x_{m_1}(k,l)x_{m_2}(k,l)|} \mathrm{e}^{j2\pi f_k \tau(m_1,m_2,\theta)} \right\}, \tag{19}$$

where the superscript $(\cdot)^*$ denotes complex conjugation, the operator $\Re\{\cdot\}$ denotes the real component of the argument, and $\tau(m_1,m_2,\theta) = \frac{\|\boldsymbol{r}_{m_2} - \boldsymbol{r}_\theta\| - \|\boldsymbol{r}_{m_1} - \boldsymbol{r}_\theta\|}{c}$ is defined as in (4).

The term $\mathrm{e}^{j2\pi f_k \tau(m_1,m_2,\theta)}$ is the inter-channel phase difference theoretically computed with the delay τ; the term $\frac{x_{m_1}(k,l)x_{m_2}^*(k,l)}{|x_{m_1}(k,l)x_{m_2}(k,l)|}$ is the inter-channel phase difference measured from x_{m_1} and x_{m_2}. The spatial likelihood γ_{TF} is high when these two inter-channel phase differences are consistent with each other. The DOA can thus be estimated as

$$\theta_{\text{TF}}(k,l) = \underset{\theta \in (-180^\circ, 180^\circ]}{\arg\max}\ \gamma_{\text{TF}}(k,l,\theta). \tag{20}$$

The localization results at individual time-frequency bins can be used to construct a spatially informed filter, which extracts the target sound coming from θ_d [34], [35]. The spatially informed filter is implemented in two steps.

In the first step, we detect the time-frequency bins that belong to the target sound, assuming that the time-frequency bins belonging to the target sound have their DOA estimates normally distributed around the mean θ_d, with variance σ_d. The detection is performed by measuring the closeness of each time-frequency bin to the target sound:

$$c_d(k,l,\theta_d) = \exp\left(-\frac{(\theta_{\text{TF}}(k,l) - \theta_d)^2}{2\sigma_d^2} \right), \tag{21}$$

where $c_d(\cdot) \in [0,1]$. The higher c_d, the higher the probability that the (k,l)-th bin is dominated by the target sound.

In the second step, we calculate a target correlation matrix, *i.e.* the correlation matrix of the target sound:

$$\boldsymbol{\Phi}_{ss}(k,l,\theta_d) = \frac{1}{L}\sum_{l=1}^{L} c_d^2(k,l,\theta_d)\boldsymbol{x}(k,l)\boldsymbol{x}^{\mathrm{H}}(k,l), \tag{22}$$

where the closeness measure $c_d(k,l,\theta_d)$ indicates the contribution of the (k,l)-th bin to the correlation matrix. Given this estimated target correlation matrix, an adaptive beamformer can be formulated easily. We use the multichannel Wiener filter [37]

$$\boldsymbol{w}_{\text{TF}}(k,l,\theta_d) = \boldsymbol{\Phi}_{xx}^{-1}(k,l)\boldsymbol{\phi}_{ss1}(k,l,\theta_d), \tag{23}$$

where $\boldsymbol{\phi}_{ss1}(k,l,\theta_d)$ is the first column of $\boldsymbol{\Phi}_{ss}(k,l,\theta_d)$, and $\boldsymbol{\Phi}_{xx}(k,l)$ is estimated directly using (12). The sound coming from θ_d is extracted as

$$y_{\text{TF}}(k,l,\theta_d) = \boldsymbol{w}_{\text{TF}}^{\mathrm{H}}(k,l,\theta_d)\boldsymbol{x}(k,l). \tag{24}$$

TABLE I
SUMMARY OF CANDIDATE ALGORITHMS FOR EGO-NOISE REDUCTION. KEY – NCM: NOISE CORRELATION MATRIX. INPUT $\boldsymbol{x}$: MICROPHONE SIGNAL; $\boldsymbol{R}$: MICROPHONE LOCATIONS; θ_d: DOA OF THE TARGET SOUND. $\mathcal{A}$: ALREADY APPLIED TO MAVS; $\mathcal{N}$: NEW FOR MAVS.

Algorithm		Abbreviation	Equation	Input	Status
Beamforming	Fixed Beamforming	`FBF`	(4)	$\boldsymbol{x}$, $\boldsymbol{R}$, θ_d	$\mathcal{A}$
	Adaptive beamforming with ideal NCM estimation	`Benchmark`	(10), (13)	$\boldsymbol{x}$	$\mathcal{N}$
	Adaptive beamforming with NCM being an identity matrix	`ABF-Identity`	(10), (15)	$\boldsymbol{x}$	$\mathcal{N}$
	Adaptive beamforming with incremental NCM estimation	`ABF-Inc`	(10), (16)	$\boldsymbol{x}$	$\mathcal{N}$
	Adaptive beamforming with NCM estimated in noise-only periods	`ABF-VAD`	(14), (16)	$\boldsymbol{x}$	$\mathcal{N}$
Blind source separation		`BSS`	(18)	$\boldsymbol{x}$	$\mathcal{A}$
Time-frequency processing		`TF`	(22), (23)	$\boldsymbol{x}$, $\boldsymbol{R}$, θ_d	$\mathcal{N}$

D. Discussion

Table I compares the beamforming, blind source separation, and time-frequency processing algorithms discussed in the previous section. The algorithms are further labelled as already applied to MAVs, $\mathcal{A}$, or new for MAVs, $\mathcal{N}$.

Compared to beamforming, BSS is more flexible as it does not require as input the locations of the microphones and the target sound, nor the VAD information. Due to the nonstationarity of the ego-noise and the low SNR, the performance of the noise correlation matrix estimation schemes (see Eqs. (13)-(16)) is limited. Estimating the correlation matrix of the MAV ego-noise is still an open problem. While fixed beamformers have already been applied to MAVs [24], [25], the use of adaptive beamformers for ego-noise reduction has not been reported yet.

BSS can extract the target sound and suppress directional ego-noise effectively by estimating the demixing matrix. However, there are several issues still unsolved when using BSS in practice. First, BSS typically works as a batch process and thus requires the acoustic mixing network to remain stationary for a certain interval, *i.e.* with physically static sound sources and microphones. Although this condition may be satisfied in some cases, *e.g.* an MAV hovering stably in the air while recording a static speaker, a dynamic mixing network is often encountered with a flying MAV. Second, the target sound is extracted into one of the M output channels with the channel index unknown. A post-processing procedure is needed to detect the target sound channel, *e.g.* by exploiting the prior knowledge of the target sound location.

The time-frequency processing approach performs ego-noise reduction by exploiting the sparsity of the acoustic signals. When applying this approach to MAVs, the locations of the microphones have to be known to estimate the DOA of the sound at each time-frequency bin. This approach also requires as input the location of the target sound in order to detect the time-frequency bins that could be used for computing the target correlation matrix. This scheme works efficiently with strong ego-noise. However, if the target sound comes from a direction close to that of an ego-noise source, the time-frequency bins belonging to the ego-noise might be erroneously detected as target sound, thus decreasing the estimation accuracy of the target correlation matrix and degrading the noise suppression performance. This is a major drawback of the time-frequency processing approach. In addition, the accuracy of local DOA estimation might drop as the reverberation density is increased. However, the fact that MAVs are mainly used in low-reverberant outdoor environments may alleviate this problem.

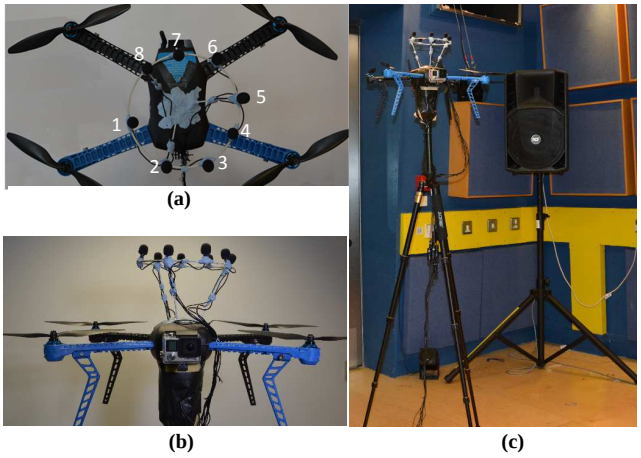


Fig. 2. The circular microphone array mounted on the MAV. (a) Top view; (b) Side view; (c) Recording environment.

IV. EXPERIMENTAL COMPARISON

A. Hardware Setup and Data

We built a hardware prototype (Fig. 2(a) and (b)) [16] composed of a circular microphone array with eight omnidirectional lapel microphones mounted on a 3DR IRIS quadcopter. The array has a 0.2 m diameter and a 0.15 m distance from the top side of the MAV. The specific mounting position of the array helps avoiding the influence of the self-generated wind blowing downwards from the propellers. The signals from the eight microphones are sampled simultaneously with a Zoom R24 recorder, at a sampling rate of 8 kHz. Fig. 2(c) depicts the recording setup in a room of size 6m×5m×3m with a reverberation time of around 200 ms. The quadcopter with microphone array is fixed on a tripod at a height of 1.8 m. A loudspeaker is placed 3 m away from the MAV and at a height of 1.3 m, playing speech signals as the target sound.

We produce two datasets, Dataset-1 and Dataset-2, while varying the speed of the motors randomly during the recording of the ego-noise. The microphone signal is generated by adding the noise and the speech at a varying input SNR from -30 dB to 5 dB, with an interval of 5 dB. Dataset-1 is produced with recorded ego-noise and simulated speech to enable a

comprehensive study. The speech is simulated with the image-source method [43] in a space of size 20m×20m×4m, with reverberation time 200 ms. The speech source is placed 10 m away, emitting a plane-wave sound at a varying DOA from 0° to 360°, with an interval of 30°. Dataset-2 is produced under a more realistic scenario with the ego-noise and the speech recorded separately. The positions of the MAV and the loudspeaker are fixed during the recording.

B. Performance Evaluation

We evaluate the performance of seven noise reduction algorithms (Table I): beamforming (`Benchmark`, `FBF`, `ABF-Identity`, `ABF-Inc` and `ABF-VAD`), BSS (`BSS`), and time-frequency processing (`TF`). We evaluate the noise reduction performance using testing signals with a 10 s duration. For all the algorithms, we set the STFT frame length as 1024, with half overlap. For `TF`, we set $\sigma_d = 10°$ in (21). For `Inc`, we set $L_t = 10$ in (16). `Benchmark` assumes the noise correlation matrix to be known. `ABF-VAD` assumes the voice activity of the target sound to be known. BSS is implemented as in [41]. For reference, we include an additional algorithm: a BSS method that assumes that the permutation ambiguities are perfectly solved by referring to the original source signals (`BSS-np`) [41].

We use SNR to measure the sound enhancement performance of a spatial filter $\boldsymbol{w}(n)$, which is a time-domain version of $\boldsymbol{w}(k,l)$. Writing the spatial filtering procedure in the time domain, it follows that

$$y(n) = \boldsymbol{w}(n) * \boldsymbol{x}(n) = \sum_{p=0}^{L_w-1} \boldsymbol{w}(p)\boldsymbol{x}(n-p)$$
$$= y_s(n) + y_v(n) = \boldsymbol{w}(n) * \boldsymbol{s}(n) + \boldsymbol{w}(n) * \boldsymbol{v}(n), \quad (25)$$

where '$*$' denotes the convolutive filtering procedure and L_w is the length of the filter $w(n)$; $y_s(n)$ and $y_v(n)$ are, respectively, the target and noise components at the output. The SNR is calculated in target-sound-active periods $\mathbb{N}_s$ as [38]

$$\text{SNR} = 10\log_{10}\frac{\sum_{n'\in\mathbb{N}_s} y_s^2(n')}{\sum_{n'\in\mathbb{N}_s} y_v^2(n')}. \quad (26)$$

We compare the SNR improvement between the input and output signals, *i.e.*

$$\text{SNR}_{\text{imp}} = \text{SNR}_{\text{out}} - \text{SNR}_{\text{in}}. \quad (27)$$

C. Results

We discuss the processing results of time-frequency processing (`TF`) in Fig. 3 for a simulated target sound (speech signal) coming from 0° and with an input SNR of -10 dB. Fig. 3(a) and Fig. 3(b) depict the time-domain waveform and the time-frequency spectrum of the input signal at one microphone. The speech signal is hardly distinguishable from the noisy background. Fig. 3(c) and Fig. 3(d), respectively, depict the time-frequency spectra of the clean ego-noise and clean speech components of the microphone signal. The time-frequency sparsity of both components can be clearly observed: the ego-noise harmonics and the speech harmonics

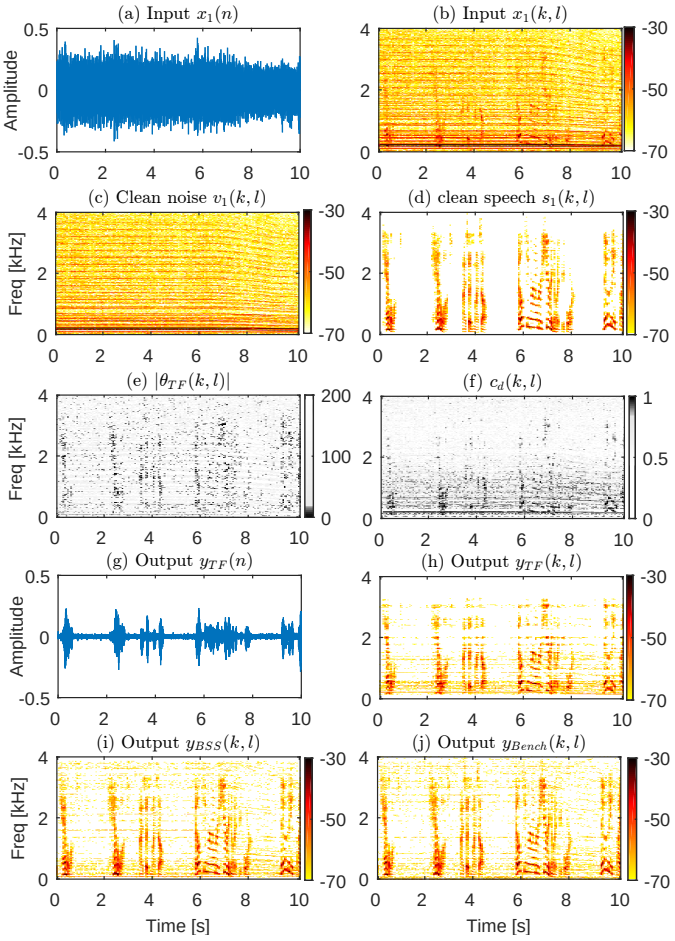


Fig. 3. Processing results with `TF`, `BSS` and `Benchmark` for a target sound with DOA 0° and input SNR -10 dB. (a)-(b): Time-domain waveform and time-frequency spectrum of the input signal; (c)-(d): Time-frequency spectra of the clean ego-noise and clean speech signals; (e) Local DOA estimation results $|\theta_{\text{TF}}(k,l)|$; (f) Contribution measure to the target sound $c_d(k,l,0)$; (g)-(h): Time-domain waveform and time-frequency spectrum of the `TF` output; (i)-(j): Time-frequency spectra of the `BSS` and `Benchmark` outputs. The output SNRs of `TF`, `BSS` and `Benchmark` are 13.9 dB, 14.4 dB and 14.5 dB, respectively.

generally occupy different time-frequency bins. Fig. 3(e) depicts the local DOA estimation results at individual time-frequency bins. For convenience of display, we plot the absolute values $|\theta_{\text{TF}}(k,l)| \in [0°, 180°]$. Most time-frequency bins that are dominated by speech components have their DOAs estimated at around 0°, distinguishing them from the background noise. Fig. 3(f) depicts the contribution measure $c_d(k,l)$ from each time-frequency bin to the computation of the target correlation matrix. Those speech-dominated time-frequency bins contribute the most to the correlation matrix. Fig. 3(g) and Fig. 3(h) depict the time-domain waveform and the time-frequency spectrum of the spatial filtering output y_{TF}, where the strong harmonic noises are almost completely removed and the speech signal can be clearly observed (SNR: 13.9 dB). For reference, Fig. 3(i) and Fig. 3(j) depict the time-frequency spectra of the output signals by `BSS` and `Benchmark`, respectively. The target sounds are well enhanced with the output SNRs being 14.4 dB and 14.5 dB, respectively.

Next, we evaluate the performance of the considered noise reduction algorithms when the target sound comes from a fixed direction with a varying input SNR from -30 dB to 5 dB, with an interval of 5 dB. For each input SNR we implement 10 realizations with different segments of noise and speech signals and calculate the averaged SNR improvement. Fig. 4(a) depicts the evaluation results for a simulated target sound coming from 0° (Dataset-1). With perfect knowledge of the noise correlation matrix, `Benchmark` outperforms all the other algorithms, with its performance invariant with respect to the varying SNR_{in}. We thus use its result as a benchmark for ego-noise reduction. The performance of `TF` and `BSS` both improves with increasing SNR_{in} for $\text{SNR}_{\text{in}} \leq$ -15 dB and then declines with increasing SNR_{in} for $\text{SNR}_{\text{in}} \geq$ -10 dB. The performance of `TF` is close to `Benchmark` for -20dB $\leq \text{SNR}_{\text{in}} \leq$ -5dB. `TF` outperforms `BSS` in almost all scenarios, especially when $\text{SNR}_{\text{in}} \leq$ -15 dB. The performance degradation of BSS in low SNRs is due to strong ego-noise, which deteriorates both ICA and permutation alignment. This is verified by comparing `BSS` and `BSS-np`, which outperforms `BSS` for all SNR_{in} especially when $\text{SNR}_{\text{in}} \leq$ -15 dB. This shows that `BSS` still suffers from permutation errors even after permutation alignment processing, and these residual permutation errors become more important when the ego-noise becomes stronger. The performance drop of `BSS-np` with decreasing SNR_{in} when $\text{SNR}_{\text{in}} \leq$ -15 dB indicates deteriorated ICA performance. By exploiting the time-frequency sparsity of the acoustic signals and the target direction, `TF` outperforms `BSS` especially in low SNRs. The performance degradation of `TF` in low SNRs is due to the broadband component of the ego-noise. With its energy uniformly distributed through the whole frequency band, the broadband noise may severely mask the target sound in low-SNR scenarios and corrupt local DOA estimation at harmonic frequencies of the target sound. For high SNR_{in}, the output SNRs of `TF` and `BSS` both rise, but with a lower rate in comparison to the increment of SNR_{in}, thus leading to declined SNR improvement with increasing SNR_{in}. The performance drop of `TF` at high SNRs also shows that the estimation error of the target correlation matrix becomes pronounced when the energy of the target sound increases with the SNR_{in}.

`TF` and `BSS` outperform the four beamforming algorithms, *i.e.* `ABF-VAD`, `ABF-Identity`, `ABF-Inc` and `FBF`, for almost all input SNRs. The poor performance of these beamforming algorithms is mainly due to their inaccurate estimation of the noise correlation matrix. Interestingly, their performance varies differently with the SNR_{in}. Assuming that a perfect VAD is available, `ABF-VAD` estimates the noise correlation matrix more accurately and hence leads to a much higher SNR improvement than the other three algorithms. However, `ABF-VAD` still performs worse than `Benchmark`, due to the nonstationarity of the ego-noise, *i.e.* the noise correlation matrix in noise-only periods is different from the one in the target-sound-active periods. The influence of this estimation error grows when the noise intensity is increased, leading to declined SNR improvement with decreasing SNR_{in}. `ABF-Identity` uses an identity matrix as the noise correlation matrix estimate. The influence of the estimation error

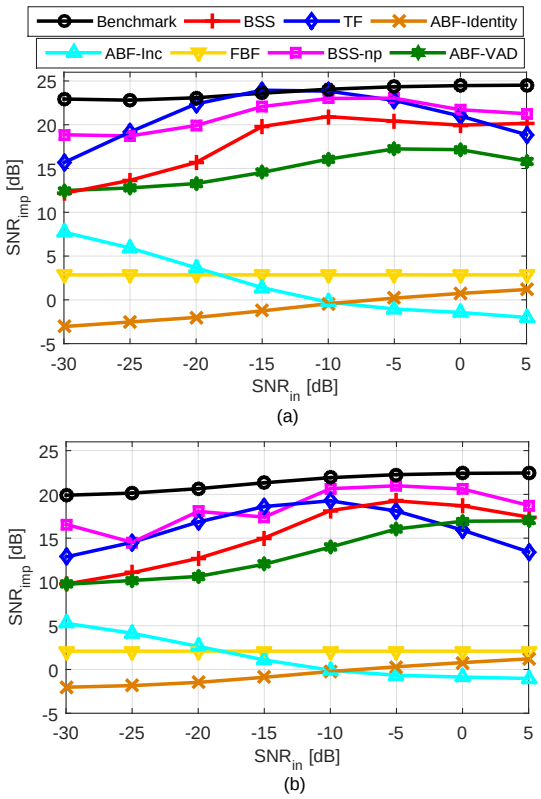


Fig. 4. SNR improvement with different noise reduction algorithms for a varying input SNR. (a) Simulated target sound with DOA 0° (Dataset-1). (b) Real-recorded target sound with DOA 160° (Dataset-2). The considered algorithms are beamforming (`Benchmark`, `FBF`, `ABF-Identity`, `ABF-Inc` and `ABF-VAD`), blind source separation (`BSS` and `BSS-np`), and time-frequency processing (`TF`). `Benchmark` assumes the noise correlation to be perfectly known. `ABF-VAD` assumes a perfect VAD. `BSS-np` assumes permutation ambiguities to be perfectly solved. A demo with the audio signals corresponding to the figures (a) and (b) is available [44].

becomes smaller when the intensity of the speech is increased, leading to increased SNR improvement with increasing SNR_{in}. `ABF-Inc` uses the microphone signal in previous frames to estimate the noise correlation matrix of the current frame. The estimation error becomes pronounced when the intensity of speech is increased, leading to decreased SNR improvement with increasing SNR_{in}. While `FBF` improves the SNR only limitedly, this improvement does not vary with the SNR_{in}.

Fig. 4(b) shows the evaluation results for a real-recorded target sound coming from 160° (Dataset-2). We can make similar observations to those made for Fig. 4(a). `Benchmark` is the best performer, followed by `BSS-np`, `TF` and `BSS`. `TF` outperforms `BSS` when $\text{SNR}_{\text{in}} \leq$ -10 dB, while `BSS` performs better when $\text{SNR}_{\text{in}} \geq$ -5 dB. `TF` and `BSS` outperform the four beamforming algorithms (`ABF-VAD`, `ABF-Identity`, `ABF-Inc` and `FBF`).

Finally, Fig. 5 shows the evaluation result for the three representative noise reduction algorithms (`Benchmark`, `BSS`, `TF`) when varying the DOA of the target sound anti-clockwise from -150° to 150°, with an interval of 30° (Dataset-1). We consider four input SNRs: -30 dB, -20 dB, 10 dB and 0 dB. For each DOA and input SNR, we implement one realization and calculate the SNR improvement. `Benchmark` produces

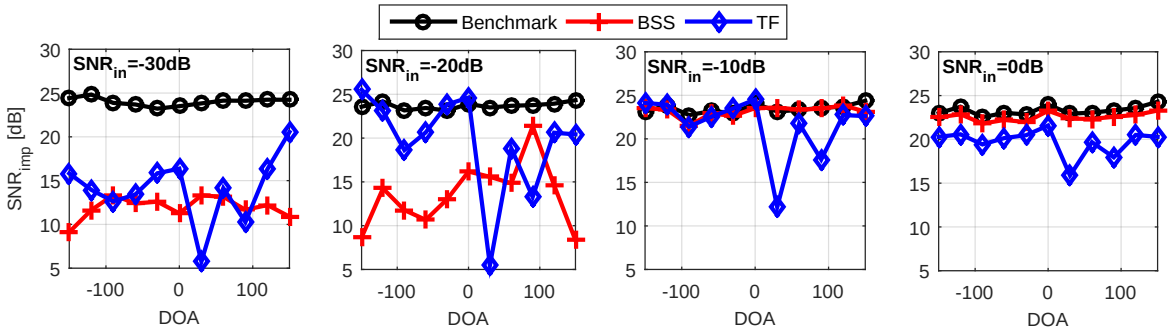


Fig. 5. SNR improvement with three noise reduction algorithms (`Benchmark`, `BSS`, `TF`) for a target sound with a varying DOA and at different input SNRs (Dataset-1).

the best result among the three algorithms and its performance does not vary with variations of DOA. The performance of `BSS` is close to that of `Benchmark` for high SNR_{in} (-10 and 0 dB) and does not change when the DOA varies. The performance of `BSS` drops significantly for low SNR_{in} (-30 and -20 dB). The performance of `TF` is sensitive to the variation of DOA. For all SNR_{in} a performance drop can be clearly observed at DOAs 30° and 90°. For low SNR_{in} (-30 and -20 dB) an additional performance drop can be observed at DOA -90°, because the ego-noise is dominant in these directions (30°, 90°, -90°). As discussed in Sec. III-D, the noise from one of these directions is detected as target sound, thus leading to an inaccurate correlation matrix and degraded noise reduction performance. When excluding these directions, `TF` performs similarly to `BSS` for high SNR_{in} (-10 and 0 dB) and outperforms BSS for low SNR_{in} (-30 and -20 dB).

V. Conclusions

We addressed the problem of acoustic sensing using multiple microphones mounted on an MAV. The main challenge is dealing with extremely low SNRs that degrade the sound recording quality significantly because of the ego-noise generated by motors and propellers. To address this problem, we proposed to use a time-frequency spatial filtering approach. We also evaluated beamforming and blind source separation algorithms that could be applied to MAV-based ego-noise reduction. Blind source separation (`BSS`) and time-frequency processing (`TF`) outperform beamforming algorithms. The biggest challenge for beamforming is the estimation of the noise correlation matrix. Due to the nonstationarity of the ego-noise, `ABF-VAD`, which assumes a perfect VAD and estimates the noise correlation matrix in noise-only periods, performs considerably worse than `Benchmark`, which assumes the noise correlation matrix to be known. As expected, the performance of `TF` degrades when the target sound arrives from a direction close to that of the ego-noise. `TF` outperforms `BSS` especially in low-SNR scenarios (*e.g.* $\leq$ -15 dB), but requires the knowledge of the DOA of the target sound, which could be obtained with an onboard camera and an object tracker [4]. Since the relative locations between the propellers and the microphones are fixed during the MAV movement, the MAV can then be intentionally rotated so that the target sound comes from a different direction from that of a propeller.

In our future work, we will investigate the effect of different array geometries placements and of the mobility of the MAV. In addition to this, we will also extend the algorithms to address scenarios with natural wind and multiple simultaneous sound sources.

References

[1] K. Daniel, S. Rohde, N. Goddemeier, and C. Wietfeld, "Cognitive agent mobility for aerial sensor networks," *IEEE Sensors J.*, vol. 11, no. 11 pp. 2671-2682, Jun. 2011.

[2] D. Floreano and R. J. Wood, "Science, technology and the future of small autonomous drones," *Nature*, vol. 521, pp. 460-466, May 2015.

[3] F. Remondino, L. Barazzetti, F. Nex, M. Scaioni, and D. Sarazzi, "UAV photogrammetry for mapping and 3D modeling - current status and future perspectives," *Int. Archives Photogrammetry, Remote Sensing Spatial Inform. Sci.*, Zurich, Switzerland, 2011, pp. 25-31.

[4] J. Pestana, J. L. Sanchez-Lopez, S. Saripalli, and P. Campoy, " Computer vision based general object following for GPS-denied multirotor unmanned vehicles," in *Proc. 2014 Amer. Control Conf.*, Portland, USA, 2014, pp. 1886-1891.

[5] F. Poiesi and A. Cavallaro, "Distributed vision-based flying cameras to film a moving target," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Hamburg, Germany, 2015, pp. 2453-2459.

[6] A. Khan, B. Rinner, and A. Cavallaro, "Multiscale observation of multiple moving targets using micro aerial vehicles," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Hamburg, Germany, 2015, pp. 4642-4649.

[7] K. Okutani, T. Yoshida, K. Nakamura, and K. Nakadai, "Outdoor auditory scene analysis using a moving microphone array embedded in a quadrocopter," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Vilamoura-Algarve, Portugal, 2012, pp. 3288-3293.

[8] M. Basiri, F. Schill, P. U. Lima, and D. Floreano, "Robust acoustic source localization of emergency signals from micro air vehicles," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Vilamoura-Algarve, Portugal, 2012, pp. 4737-4742.

[9] S. Uemura, O. Sugiyama, R. Kojima, and K. Nakadai, "Outdoor acoustic event identification using sound source separation and deep learning with a quadrotor-embedded microphone array," in *Proc. Int. Conf. Adv. Mechatronics*, Tokyo, Japan, 2015, pp. 329-330.

[10] S. Lana, K. Takahashi, and T. Kinoshita, "Consensus-based sound source localization using a swarm of micro-quadrocopters," in *Proc. Robot. Soc. Japan*, Tokyo, Japan, 2015, pp. 1-4.

[11] T. Latif, E. Whitmire, T. Novak, and A. Bozkurt, "Sound localization sensors for search and rescue biobots," *IEEE Sensors J.*, vol. 16, no. 10, pp. 3444-3453, May 2016.

[12] M. Basiri, F. Schill, P. Lima, and D. Floreano, "On-board relative bearing estimation for teams of drones using sound,", *IEEE Robot. Autom. Lett.*, vol. 1, no. 2, pp. 820-827, 2016.

[13] S. Yoon, S. Park, Y. Eom, and S. Yoo, "Advanced sound capturing method with adaptive noise reduction system for broadcasting multicopters," in *Proc. IEEE Int. Conf. Consum. Electron.*, Las Vegas, USA, 2015, pp. 26-29.

[14] J. Klapel, *Acoustic Measurements with a Quadcopter: Embedded System Implementations for Recording Audio from Above*, Master Thesis, Norwegian University of Science and Technology, 2014.

[15] G. Sinibaldi and L. Marino, "Experimental analysis on the noise of propellers for small UAV," *Appl. Acoust.*, vol. 74, no. 1, pp. 79-88, Jan. 2015.

[16] L. Wang and A. Cavallaro, "Ear in the sky: Ego-noise reduction for auditory micro aerial vehicles", in *Proc. Int. Conf. Adv. Video Signal-Based Surveillance*, Colorado Springs, USA, 2016, pp. 1-7.

[17] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*, Berlin, Germany: Springer-Verlag, 2008.

[18] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, "Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity using multiple microphones," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 18-30, Feb. 2015.

[19] L. Wang, T. K. Hon, J. D. Reiss, and A. Cavallaro, "An iterative approach to source counting and localization using two distant microphones,", *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 6, pp. 1079-1093, Jun. 2016.

[20] L. Wang, J. D. Reiss, and A. Cavallaro, "Over-determined source separation and localization using distributed microphones,", *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 9, pp. 1569-1584, Sep. 2016.

[21] H. G. Okuno and K. Nakadai, "Robot audition: Its rise and perspectives," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Brisbane, Australia, 2015, pp. 5610-5614.

[22] S. Argentieri, P. Danes, and P. Soueres, "A survey on sound source localization in robotics: From binaural to array processing methods," *Computer Speech Lang.* vol. 34, no. 1, pp. 87-112, 2015.

[23] P. Marmaroli, X. Falourd, and H. Lissek, "A UAV motor denoising technique to improve localization of surrounding noisy aircrafts: proof of concept for anti-collision systems," in *Proc. Acoust.*, 2012, pp. 1-6.

[24] T. Ishiki and M. Kumon, "A microphone array configuration for an auditory quadrotor helicopter system," in *Proc. IEEE Int. Symp. Safety, Security, Rescue Robot.*, Toyako-cho, Japan, 2014, pp. 1-6.

[25] T. Ishiki and M. Kumon, "Design model of microphone arrays for multirotor helicopters," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Hamburg, Germany, 2015, pp. 6143-6148.

[26] S. Yoon, S. Park, and S. Yoo, "Two-stage adaptive noise reduction system for broadcasting multicopters," in *Proc. IEEE Int. Conf. Consum. Electron.*, Las Vegas, USA, 2016, pp. 219-222.

[27] R. P. Fernandes, E. C. Santos, A. L. L. Ramos, and J. A. Apolinario Jr., "A first approach to signal enhancement for quadcopters using piezoelectric sensors," in *Proc. Int. Conf. Transformative Sci. Eng. Business Social Innovation*, Fort Worth, USA, 2015, pp. 536-541.

[28] K. Furukawa, K. Okutani, K. Nagira, T. Otsuka, K. Itoyama, K. Nakadai, and H. G. Okuno, "Noise correlation matrix estimation for improving sound source localization by multirotor UAV," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Tokyo, Japan, 2013, pp. 3943-3948.

[29] G. Ince, K. Nakamura, F. Asano, H. Nakajima, and K. Nakadai, "Assessment of general applicability of ego noise estimation," in *Proc. IEEE Int. Conf. Robot. Autom.*, Shanghai, China, 2011, pp. 3517-3522.

[30] G. Ince, K. Nakadai, and K. Nakamura, "Online learning for template-based multi-channel ego noise estimation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Vilamoura-Algarve, Portugal, 2012, pp. 3282-3287.

[31] T. Tezuka, T. Yoshida, and K. Nakadai, "Ego-motion noise suppression for robots based on semi-blind infinite non-negative matrix factorization," in *Proc. IEEE Int. Conf. Robot. Autom.*, Hong Kong, China, 2014, pp. 6293-6298.

[32] V. Tourbabin, H. Barfuss, B. Rafaely, and W. Kellermann, "Enhanced robot audition by dynamic acoustic sensing in moving humanoids," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Brisbane, Australia, 2015.

[33] S. Makino, T. W. Lee, and H. Sawada, Eds. *Blind speech separation*, Berlin, Germany: Springer-Verlag, 2007.

[34] K. Kowalczyk, O. Thiergart, M. Taseska, G. Del Galdo, V. Pulkki, and E. A. P. Habets, "Parametric spatial sound processing: a flexible and efficient solution to sound scene acquisition, modification, and reproduction," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 31-42, Mar. 2015.

[35] O. Thiergart, M. Taseska, and E. A. P. Habets, "An informed parametric spatial filter based on instantaneous direction-of-arrival estimates," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 2182-2196, Dec. 2014.

[36] F. Asano, S. Hayamizu, T. Yamada, and S. Nakamura, "Speech enhancement based on the subspace method," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 5, pp. 497-507, Sep. 2000.

[37] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Trans. Signal Process.*, vol. 50, no. 9, pp. 2230-2244, Sep. 2002.

[38] L. Wang, T. Gerkmann, and S. Doclo, "Noise power spectral density estimation using MaxNSR blocking matrix," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 9, pp. 1493-1508, Sep. 2015.

[39] T. Ohata, K. Nakamura, T. Mizumoto, T. Taiki, and L. Nakadai, "Improvement in outdoor sound source detection using a quadrotor-embedded microphone array," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Chicago, USA, 2014, pp. 1902-1907.

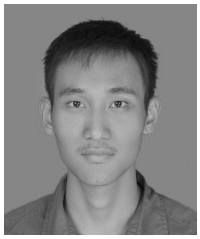
[40] L. Wang, H. Ding, and F. Yin, "A region-growing permutation alignment approach in frequency-domain blind source separation of speech mixtures," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 3, pp. 549-557, Mar. 2011.

[41] L. Wang, "Multi-band multi-centroid clustering based permutation alignment for frequency-domain blind speech separation," *Digit. Signal Process.*, vol. 31, pp. 79-92, 2014.

[42] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, New York, USA: John Wiley & Sons, 2004.

[43] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943-950, 1979.

[44] http://www.eecs.qmul.ac.uk/∼andrea/auditory-mav.html



Lin Wang received the B.S. degree in electronic engineering from Tianjin University, China, in 2003; and the Ph.D degree in signal processing from Dalian University of Technology, China, in 2010. From 2011 to 2013, he was an Alexander von Humboldt Fellow at the University of Oldenburg, Germany. Since 2014, he has been a postdoctoral researcher in the Centre for Intelligent Sensing at Queen Mary University of London. His research interests include video and audio compression, microphone array, blind source separation, and 3D audio processing.

Andrea Cavallaro received the Ph.D. degree in electrical engineering from Swiss Federal Institute of Technology, Lausanne, Switzerland, in 2002. He was a Research Fellow with British Telecommunications in 2004. He is a Professor of Multimedia Signal Processing and the Director of the Centre for Intelligent Sensing at Queen Mary University of London. He has authored more than 150 journal and conference papers, one monograph on Video Tracking (Wiley, 2011), and three edited books, Multi-Camera Networks (Elsevier, 2009), Analysis, Retrieval and Delivery of Multimedia Content (Springer, 2012), and Intelligent Multimedia Surveillance (Springer, 2013). Prof. Cavallaro is Senior Area Editor of IEEE TRANSACTIONS ON IMAGE PROCESSING and Associate Editor of the IEEE MultiMedia Magazine. He is an elected member of the IEEE Image, Video, and Multidimensional Signal Processing Technical Committee, and is the Chair of its Awards Committee, and an elected member of the IEEE Circuits and Systems Society Visual Communications and Signal Processing Technical Committee. He is a former elected member of the IEEE Signal Processing Society Multimedia Signal Processing Technical Committee, Associate Editor of IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON SIGNAL PROCESSING and IEEE TRANSACTIONS ON IMAGE PROCESSING, and Associate Editor and Area Editor of IEEE Signal Processing Magazine, and Guest Editor of eleven special issues of international journals. He was General Chair for IEEE/ACM ICDSC 2009, BMVC 2009, M2SFA2 2008, SSPE 2007, and IEEE AVSS 2007. He was Technical Program Chair of IEEE AVSS 2011, EUSIPCO 2008, and WIAMIS 2010. He received the Royal Academy of Engineering Teaching Prize in 2007, three Student Paper Awards at IEEE ICASSP in 2005, 2007, and 2009, respectively, and the Best Paper Award at IEEE AVSS 2009.